Inverse network sampling to explore online brand allegiance

Inverse Network Sampling to Explore On-line Brand Allegiance*

Peter Grindrod† Desmond J. Higham‡ Peter Laflin§
Amanda Otley¶ Jonathan A. Ward, ‖

January 15, 2016

Abstract

Within the online media universe there are many underlying communities. These may be defined, for example, through politics, location, health, occupation, extracurricular interests or retail habits. Government departments, charities and commercial organisations can benefit greatly from insights about the structure of these communities; the move to customer-centered practices requires knowledge of the customer base. Motivated by this issue, we address the fundamental question of whether a subnetwork looks like a collection of individuals who have effectively been picked at random from the whole, or instead forms a distinctive community with a new, discernible structure. In the former case, to spread a message to the intended user base it may be best to use traditional broadcast media (TV, billboard), whereas in the latter case a more targeted approach could be more effective. In this work, we therefore formalize a concept of testing for substructure and apply it to social interaction data. First, we develop a statistical test to determine whether a given subnetwork (induced subgraph) is likely to have been generated by sampling nodes from the full network uniformly at random. This tackles an interesting inverse alternative to the more widely studied "forward" problem. We then apply the test to a Twitter reciprocated mentions network where a range of brand name based subnetworks are created via tweet content. We correlate the computed results against the independent views of sixteen digital marketing professionals. We conclude that there is great potential for social media based analytics to quantify, compare and interpret on-line brand allegiances systematically, in real time and at large scale.

Keywords: generating function, mentions, networks, p-value, random graph, sampling, statistics, Twitter.

*Submitted to the European Journal of Applied Mathematics **Special Issue on Networks**.
†Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK
‡Department of Mathematics and Statistics, University of Strathclyde, Glasgow, G1 1XH, UK
§Bloom Agency, Marshalls Mill, Marshall Street, Leeds, LS11 9YJ, UK
¶Bloom Agency, Marshalls Mill, Marshall Street, Leeds, LS11 9YJ, UK
‖School of Mathematics, University of Leeds, Leeds, LS2 9JT, UK

1 Motivation

Online digital footprints are a rich source of information about human behaviour [21]. The scale and timeliness of such data sets can provide new insights of great social and commercial value [1, 9, 11, 16]. Accurate, microscale data allows us to develop and test hypotheses about the way we interact [21], and organisations now exploit social media as a means to

- discover and target significant individuals or timepoints [2, 15, 17],
- gather or distribute key messages [3, 4, 17],
- gain higher-level understanding about the nature of relevant communities [14, 18, 24, 27], and
- produce actionable summaries of the views and opinions of a general population [5, 9].

Motivated by these developments, in this work we indentify a relevant inverse problem in statistical network theory, develop a simple algorithm, and validate its performance against the views of social media experts on a new data set that we make publicly available.

We focus on brand allegiance, which is now of wide interest in the context of on-line user-communities. In [28] an empirical study of Facebook data demonstrated the existence and importance of brand communities. Modeling around an opinion survey in [20] led to the conclusion that (page 1763) "brand communities established on social media enhance feelings of community among members and contribute to creating value for both members and the company."A study in [6] argued that consumer engagement in a virtual brand community "enhances loyalty and satisfaction, empowerment, connection, emotional bonding, trust and commitment." Related recent work in [19] also found that social media-based brand communities enhance brand trust and loyalty. The authors of [7] defined *reciprocity* in a virtual community (VC) as "voluntary and discretionary behaviors in terms of giving help to not only those who help the giver but also other members in the VC who need help and who would provide assistance on request." Studying the cosmetic message board of a popular Chinese women's website, the authors concluded that (page 1038) "consumers who establish strong social ties, experience fun, and exhibit reciprocity likely collaborate with other consumers to purchase online."

This evidence suggests that virtual brand communities are of social and commercial relevance. However, since there are many different types of brand, and many differences between the off-line and on-line worlds [8, 10], it is clearly of interest to understand how these communities may vary. In particular, an organisation would benefit from knowledge of the global structure of its user-base, not least in order to devise strategies to reach out effectively to these customers. In this interdisciplinary work, which has been conducted jointly between mathematical scientists and colleagues working in a commercial digital marketing agency, we therefore formulate and address a simple, useful and quantifiable question:

> Does the observed community appear to be a randomly chosen subcollection of individuals?

In addition to increasing our basic understanding of on-line human behaviour, quantifying an answer to this question gives us actionable insights about the effectiveness of social media channels for broadcasting messages and engaging communities.

The rest of the article is organised a follows. In section 2 we discuss previous work in network analysis aimed at the "forward" problem of applying a sampling strategy to a large network and observing the resulting subnetwork. Section 3 sets up the mathematical notation and covers the relevant background on degree distributions. In section 4 we then formalize the new inverse problem and in section 5 a statistical test is devised. Section 6 motivates the application of the test in a social media setting. Section 7 describes an experiment on Twitter data, where we apply the new methodology and interpret the results alongside the views of social media professionals. A brief summary is given in Section 8.

2 Network Sampling

Many authors have commented on the fact that network samples can produce misleading pictures—that is, given a large network, a "random" collection of its nodes and edges may have a very different structure. This idea was was formalized in the seminal work [26], where it was shown analytically that selecting nodes at random from a network with a scale-free degree distribution can produce subnetworks without this characteristic pattern. Such a result has implications in many fields where the network that we observe (such as a set of experimentally tested protein-protein interactions, or a list of sexual partners obtained through individual history mapping) serves as a proxy for a larger network that is out of reach. Following on from [26], related empirical work in [23] compared different sampling strategies, such as selection of nodes at random, selection of nodes at random plus all their neighbours, selection of nodes by a random walk and selection of edges at random. These were quantified through their ability to recover a range of graph properties from the full network, including in-degree and out-degree distribution. Related empirical work in [22] considered degree distribution, path length, betweenness centrality, assortativity and clustering. From a slightly different perspective the authors in [25] took the view that sampling bias may have a beneficial effect. Based on an empirical study comparing several sampling approaches on real data sets, they concluded that, in some circumstances, discrepancies arising from a sampling strategy may be an asset; for example allowing important nodes to be located efficiently. Recent work in [13] has also highlighted the benefit of targeted sampling for efficient identification of important nodes. Here the use of neighbours of randomly chosen nodes is particularly effective, due to the so-called Friendship Paradox [12].

The work discussed above is focussed on the forward problem: what are the properties of the sampled subnetwork, given the full network and a sampling procedure? Our work differs by addressing an associated inverse problem: given a network and an observed subnetwork, what can we deduce about the manner in which the full network

was sampled? More precisely, we focus on a specific question: was the subnetwork obtained by choosing nodes of the full network uniformly and independently at random? As discussed in section 1, and elaborated upon in section 6, in the context of brand-specific communities within a larger social media environment, this question may be formulated as: does the virtual community related to brand X look like a random sample of independently chosen individuals?

3 Technical Background

Given an undirected, unweighted graph, $\mathcal{G}$, over a large number of vertices, N, we may compute the degree distribution, $\{P_k | k = 0, .., N-1\}$. Here, P_k records the proportion of nodes with degree k. The associated probability generating function has the form

$$G(x) = \sum_{k=0}^{N-1} P_k x^k. \tag{1}$$

The mean vertex degree and variance are given from (1) by

$$z_{\mathcal{G}} := \langle k \rangle = \sum_{k=0}^{N-1} k P_k = G'(1),$$

and

$$\sigma^2_{\mathcal{G}} := \langle (k - z_{\mathcal{G}})^2 \rangle = \sum_{k=0}^{N-1} (k - z_{\mathcal{G}})^2 P_k = G''(1) + z_{\mathcal{G}} - z^2_{\mathcal{G}},$$

respectively, where prime denotes differentiation.

Suppose we create a subgraph $\mathcal{H}$ of $\mathcal{G}$ by drawing vertices independently and identically from those of $\mathcal{G}$, each with probability α, and include any edges inherited from $\mathcal{G}$. If a vertex of degree k in $\mathcal{G}$ is drawn then it has degree $k' \leq k$ in $\mathcal{H}$ with probability

$$P(k' \text{ in } \mathcal{H} | k \text{ in } \mathcal{G}) = \alpha^{k'} (1-\alpha)^{k-k'} \begin{pmatrix} k \\ k' \end{pmatrix}.$$

Let $H(x)$ denote the generating function for $\mathcal{H}$. It follows that

$$H(x) = G(1 - \alpha + \alpha x), \tag{2}$$

since a vertex with degree k in $\mathcal{G}$ that is selected for $\mathcal{H}$ contributes the term $(1-\alpha+\alpha x)^k$ to the sum for H. The relation (2) was derived in [26] in order to show that randomly sampled subnetworks cannot generally possess the same properties as the full networks (in that particular case a scale-free degree distribution). However if $G(x)$ happens to be in the form

$$G(x) = F(1 + \mu(x-1))$$

for some real μ and some given function F, for which $F(1) = 1$, then (2) implies

$$H(x) = F(1 + \mu\alpha(x-1)).$$

So the well known class of negative binomial distributions (and hence Poisson distributions) are invariant under independent vertex sampling (only the coefficient of $(x-1)$ changes), again, see [26].

4 An Inverse Problem

It is fruitful to think of using (2) in the opposite direction. If we observe $\mathcal{H}$, as a sampled graph, and estimate the Maclaurin series defining the generating function, H, then what may be said about G and α? We have

$$G(x) = H(1 - (1 - x)/\alpha), \tag{3}$$

so the degree distribution for $\mathcal{G}$, defined by the $P_k \in [0, 1]$ in (1), are related to the derivatives of H evaluated at $1 - 1/\alpha < 0$:

$$P_k = \frac{H^{(k)}(1 - 1/\alpha)}{\alpha^k}.$$

Clearly we require H and all derivatives of H to remain positive in the interval $(1 - 1/\alpha, 1]$.

5 Testing Given Subgraphs

Using (2) we may estimate the mean and standard deviation for the vertex degree distribution of a subgraph $\mathcal{H}$ under the hypothesis of random independent selection of vertices from $\mathcal{G}$:

$$z_{\mathcal{H}} = H'(1) = \alpha z_{\mathcal{G}},$$

and

$$\sigma^2_{\mathcal{H}} = \alpha^2 G''(1) + \alpha z_{\mathcal{G}} - \alpha^2 z^2_{\mathcal{G}},$$

so that

$$\sigma^2_{\mathcal{H}} = \alpha^2(\sigma^2_{\mathcal{G}} - z_{\mathcal{G}} + z^2_{\mathcal{G}}) + \alpha z_{\mathcal{G}} - \alpha^2 z^2_{\mathcal{G}} = \alpha^2 \sigma^2_{\mathcal{G}} + \alpha(1 - \alpha) z_{\mathcal{G}}. \tag{4}$$

Moreover the central limit theorem tells us that the observed mean degree for vertices in such a subgraph, $\mathcal{H}$, having n vertices, is distributed approximately normally about $z_{\mathcal{H}}$ with variance $\sigma^2_{\mathcal{H}}/n$.

Now suppose we are given $\tilde{\mathcal{H}}$, some subgraph of $\mathcal{G}$, with $n < N$ vertices, which is selected by some criterion or other (whether known to us or not). Let us set $\alpha = n/N$ to create the suitable null hypothesis that $\tilde{\mathcal{H}}$ was randomly drawn from $\mathcal{G}$, and thus calculate $z_{\mathcal{H}}$ and $\sigma_{\mathcal{H}}$, for randomly drawn subnetworks of the correct size. Suppose that the mean degree observed for $\tilde{\mathcal{H}}$, denoted by $z_{\tilde{\mathcal{H}}}$, is bigger than $z_{\mathcal{H}} = \alpha z_{\mathcal{G}}$. Then the probability that a graph drawn under the above random hypothesis has a mean degree greater than or equal to $z_{\tilde{\mathcal{H}}}$ is given by

$$Q(z_{\tilde{\mathcal{H}}}, n) = \frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{n}\,(z_{\tilde{\mathcal{H}}}/\alpha - z_{\mathcal{G}})}{\sqrt{2}\sqrt{\sigma^2_{\mathcal{G}} + (1/\alpha - 1) z_{\mathcal{G}}}}\right). \tag{5}$$

Hence we have a (one sided) p-value. If this value is very small then the null hypothesis is very unlikely and so the criterion used to select $\tilde{\mathcal{H}}$ cannot be uncorrelated with the structure that is observed in $\mathcal{G}$.

For example, suppose we have $\mathcal{G}$ with $P_k \propto w_k(1+k)^{-3.1}$ where each w_k is chosen independently from a uniform distribution over [0,1]; and let us take $N = 1000$. So $\mathcal{G}$ is a scale free network (note, the P_k's are normalised to sum to one). Then we may plot the contours of the Q given by (5) in the $(n, z_{\tilde{\mathcal{H}}})$-plane, see Figure 1. Thus for any given $\tilde{\mathcal{H}}$ we obtain a p-value under the null hypothesis that its vertices were drawn randomly from those of $\mathcal{G}$.

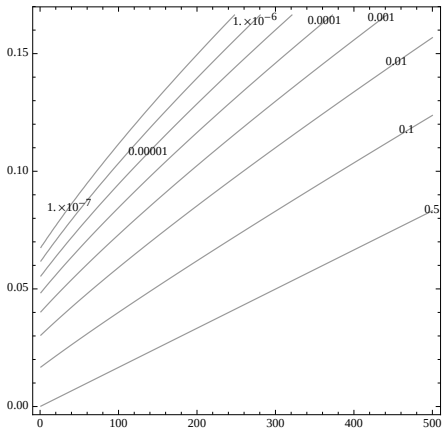

Figure 1: Contours of Q given by (5), with n on the horizontal axis and $z_{\tilde{\mathcal{H}}}$ on the vertical axis, representing a p-value for any observed subgraphs located there ($N = 1000$ and $P_k \sim (1+k)^{-3.1}$).

To test this idea, we took the network described in section 7, for which there are $N = 30,191$ nodes. Here we have mean vertex degree $z_{\mathcal{G}} = 1.82491$ and variance $\sigma^2_{\mathcal{G}} = 6.36821$. We sampled induced subgraphs $\mathcal{H}$ by choosing 1000 vertices uniformly at random, so $\alpha = 0.033122$, in each case computing the mean vertex degree $z_{\mathcal{H}}$. Using 500 separate samples for $\mathcal{H}$, we obtained a sample mean for $z_{\mathcal{H}}$ of 0.058088 compared with $\alpha z_{\mathcal{G}} = 0.6044$. The observed variance for the distribution of the 500 $z_{\mathcal{H}}$'s was 0.0001397. From the Central Limit Theorem, we predict this value to agree with $\sigma^2_{\mathcal{H}}/n$. Direct calculation of $\sigma^2_{\mathcal{H}}$ using (4) produced 0.0001309, which closely matches this prediction.

6 Subgraph Non-Randomness and Social Media

We focus now on the case where the graph $\mathcal{G}$ is generated via an online social network platform. Users form the vertices of the network and edges represent pairwise interactions generated via directed messaging. Now suppose that, independently of the link structure, we select subsets of the vertices according to whether they display any interest in certain *topics*, indicated by the appearance of key works or phrases within the messages. For example, a topic may correspond to a consumer brand, a news item, a political opinion, a health issue, an event, a TV show or a celebrity. Each such topic induces a subgraph, $\tilde{\mathcal{H}}$. We may observe its size, n, and mean degree, $z_{\tilde{\mathcal{H}}}$, and use (5)

to calculate its p-value. It may be argued that for ubiquitous topics, such as asthma or Ford Motor Company we should expect to find values greater than 10^{-1}, indicating that the topic-based subnetwork is not dissimilar to one whose vertices are drawn at random from $\mathcal{G}$. For less pervasive topics, perhaps Lexus cars or leukaemia, we may expect to find p-values that are vanishingly small, indicating that the topic-based subnetwork has an anomalous amount of structure that would not be present in a random sample.

Of course, we cannot infer causality in this way. Lexus devotees may have other lifestyle choices in common, and hence may form links for a variety of reasons, not simply through a shared automotive preference. Hence the *topic* may not *cause* the heightened community structure—but will *reflect* it. On the other hand, individuals suffering from leukaemia, or their family members and friends, may meet physically at special clinics, seek each other online for advice and support, or take part in leukaemia-based events. In such cases, a common leukaemia interest is likely to *drive* enhanced social connectivity.

Irrespective of causality, consider now whether social media would be a good channel for engagement by third parties, such as marketing companies, government/public health advisors or journalists. If the p-value for a topic is relatively high (say $> 10^{-1}$) then any attempt to channel information over the same social media platform is nothing more than broadcasting: you may advertise on buses to reach this random subset more effectively. On the other hand for a very low p-value, the topic-based community has an anomalous amount of structure, far from that of a random selection, and so targeted information can be passed around efficiently within the structured target sub-population.

This idea was first discussed in [14], where the possible justification of investments into targeted online marketing was considered, albeit with a p-value determined experimentally by (re-)sampling. Here we have removed the need for such sampling (which is computationally infeasible for the case of very small p-values) through the use of (5), fixing on the mean degree as our comparative measure of community structure.

7 Experimental Results

In order to apply our statistical test to real data, we constructed a reciprocated Twitter mentions network. Here an undirected link is inserted between nodes i and j if and only if i was observed to mention j at least once and, additionally, j was observed to mention i at least once. We used a full set of Tweets that were geolocated in any one of ten UK cities: Birmingham, Bristol, Cardiff, Edinburgh, Glasgow, Leeds, London, Manchester, Nottingham and Sheffield. The data was collected over a period of ten days, from 1st October to 10th October, 2014. This produced an undirected network of $N = 30,191$ nodes, with $27,548$ edges. Within this network, we extracted nodes that had created tweets containing keywords relating to the following fifteen issues:

- two UK football teams: Leeds United and Cardiff City,
- four high-profile retailers: John Lewis, Marks & Spencer (M&S), Morrisons and Tesco,

- two car manufacturers: Jaguar and Vauxhall,
- seven recognisable product lines: Corona (beer), Colgate, Diet Coke, Marmite, Monster (energy drink), Pot Noodle and Stella Artois.

In Table 1 we list the subnetworks in order of increasing p-value; so the entries at the top of the list are least likely to have arisen via independent uniform random sampling of nodes. The table also shows the number of nodes and edges in each subnetwork, the mean degree that would arise from uniform sampling ($\alpha z_{\mathcal{G}}$) and the observed mean degree.

We see from Table 1 that all but one of the subnetworks has a higher mean degree than that predicted by random sampling; that is, $z_{\tilde{\mathcal{H}}} > \alpha z_{\mathcal{G}}$. If we use the standard cut-off of 0.05 then all subnetworks are significantly different from a random sample (note that $\log_{10}(0.05) \approx -1.3$). However, we also see a wide spread of p-values. In terms of ordering, the test clearly places the two football team subnetworks at the top of the list, with Leeds United ahead of Cardiff City. This agrees with the experience of some of the authors (who work in the social media analytics team of a Leeds-based company)—the Leeds United supporting community is regarded as more tightly packed than that of Cardiff City.

As a follow-on experiment to give more insight into these results, we enlisted the help of sixteen professionals from a digital social media agency, and asked them to rank the fifteen brands. We emphasize that these colleagues were not given access to any Twitter data, instead they were asked to rely on their knowledge of the brand, gained through their professional experience. More precisely, they were sent an email with the following instructions:

> *What I'd like you to do is to reply to this email and put these "things" in order based on how likely it is that a group of people would come together and talk about that "thing." Some are more obvious than others and please don't let your like/dislike for a "thing" influence your choice. I'm interested in whether people would/could talk about these "things" or whether they wouldn't.*

Table 2 shows the rankings returned by each participant. Here, we have ordered the brands according to their overall rank across the sixteen responses.

We see in Table 2 that the two football teams are consistently ranked highly, in agreement with the statistical tests. Generally, the social media team, in common with the social media ranking in Table 1, placed "everyday" products below the retail outlets.

In an attempt to compare directly the two sets of results, Table 3 shows the pairwise Kendall tau correlation for the social media professionals' rankings and the statistically produced ranking. We note a strong level of agreement, with strictly positive correlation between the social media ranking and each human ranking (final column). As a summary, Table 4 shows, for each ranking, the sum of the pairwise correlations across all other rankings. The left hand column of Table 4 corresponds to Kendall tau correlation, and the right hand column to Spearman rho. We see that in this specific, well-defined, sense, the automated ranking cannot be distinguished from that of a social media professional. We view this level of consistency as extremely promising,

given that (a) the statistical ranking is based on physical data from real, recent, interactions, whereas the humans made decisions based on background knowledge, and (b) the instructions given to the humans were necessarily brief and subject to interpretation. Furthermore, the wide variation of views between the sixteen colleagues makes it clear that there is no absolute ground truth against which to judge an algorithm.

Feedback on these results from the social media professionals emphasized to us that the new methodology investigates how people are *actually* talking about these brands and the connectedness of the conversations, whereas typical marketeers must rely on their *intuition* concerning how these brands are talked about. Given evidence that these two sources of information are broadly compatible, the differences then become fascinating.

The most striking discrepancies involve Vauxhall and Marmite. The social media information suggests that they are talked about much more than we may think. Marmite is anecdotally cited as a "love-hate" product that polarises opinion. Vauxhall have a well-defined community around their sponsorship of football[1], and are likely to be benefiting from that involvement. On this basis, our social media colleagues felt that the computational analysis added value to their preconceptions. On a similar note, the "high-end" retailers John Lewis and M&S were given relatively low ranking by the algorithm. The social media professionals were interested to find that, based on their perceived value of these two brands, people talk about John Lewis and M&S much less than they had expected.

8 Discussion

The main novelties in this work were (a) motivating and defining a new inverse network sampling problem, (b) developing a new algorithm to address the problem, (c) applying the algorithm to new social media data, which will be made publicly available, and (d) validating the results against independent expert knowledge. Among the advantages of this type of automated data-driven approach are that

- large-scale data sets can be summarized and compared systematically in a manner that is easy to explain,
- computations can be updated and monitored in real time in order to monitor dynamic changes.

The research was developed and tested in the context of virtual brand communities, where it has the potential to help us understand on-line behaviour and may therefore lead to improved products and services in customer-facing industries. However, we note that the methodology can be applied in any network setting where there are well-defined subcommunities whose structure is of interest.

Acknowledgements
We are grateful to Bloom Agency, Leeds, UK, for giving us access to the social media experts who contributed to this study. The Twitter data will be made available upon

[1] `https://www.vauxhallfootball.co.uk/england/Fanbase`

$\log_{10}$ p-val.	issue	nodes	edges	$\alpha z_{\mathcal{G}}$	$z_{\tilde{\mathcal{H}}}$
-22267	Leeds United	1377	1866	0.083	2.7
-2323	Cardiff City	289	120	0.017	0.83
-447	Vauxhall	238	44	0.014	0.37
-352	Marmite	94	15	0.0057	0.32
-292	Tesco	881	153	0.053	0.35
-141	Jaguar	79	8	0.0048	0.20
-134	John Lewis	167	17	0.010	0.20
-74	Diet Coke	55	4	0.0033	0.15
-70	Stella	110	8	0.0066	0.15
-67	M&S	99	7	0.0060	0.14
-57	Colgate	16	1	0.00097	0.13
-57	Corona	16	1	0.00097	0.13
-54	Pot Noodle	64	4	0.0039	0.13
-50	Monster	17	1	0.0010	0.12
-28	Morrisons	183	9	0.011	0.10

Table 1: Results for reciprocated Twitter mention subnetworks. Subnetworks are listed in order of increasing p-value; those at the top are least likely to arise from sampling the full network uniformly at random. Column 1: $\log_{10}$ of p-value. Column 2: issue defining the subnetwork. Column 3: number of nodes. Column 4: number of edges. Column 5: mean subnetwork degree that would arise from uniform sampling. Column 6: observed mean subnetwork degree.

Leeds United	3	1	1	1	6	1	1	1	1	1	2	1	1	3	2	8
Cardiff City	4	2	2	2	5	2	2	2	2	2	1	2	3	4	5	5
John Lewis	1	6	7	13	1	7	4	4	9	5	4	7	10	7	3	1
Jaguar	9	3	5	7	3	5	8	3	3	13	7	6	14	2	13	14
M&S	2	9	6	12	8	8	5	7	10	6	5	8	11	8	8	2
Stella	6	4	3	4	10	4	9	13	5	8	6	13	7	10	11	4
Diet Coke	15	11	10	5	2	9	3	8	7	4	9	15	2	13	7	6
Tesco	13	7	4	11	7	10	11	5	14	7	12	14	4	5	1	3
Corona	8	5	11	3	9	3	10	14	6	14	3	5	8	11	14	7
Vauxhall	14	8	9	8	11	6	13	9	4	3	14	3	13	1	10	10
Morrisons	12	10	8	10	14	11	12	10	11	9	11	11	9	6	4	9
Marmite	5	15	12	14	4	13	6	12	8	11	10	9	12	14	6	11
Pot Noodle	11	12	13	9	13	15	7	6	13	10	13	12	6	15	9	12
Monster	10	13	15	6	12	14	14	11	12	12	8	10	5	9	12	15
Colgate	7	14	14	15	15	12	15	15	15	15	15	4	15	12	15	13

Table 2: Rankings from humans, independently of the social media data. Each column shows the independent ranking of one social media professional, in response to the request shown in the text. The topics are ordered according to their average rank across these experts.

Humans															
2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	p-val
0.18	0.22	0.03	0.20	0.24	0.37	0.10	0.18	0.01	0.49	0.41	**-0.07**	0.01	0.07	0.26	0.24
	0.70	0.47	0.30	0.75	0.40	0.50	0.62	0.37	0.50	0.31	0.22	0.49	0.16	0.28	0.37
		0.31	0.30	0.60	0.39	0.50	0.47	0.45	0.39	0.16	0.22	0.45	0.39	0.43	0.49
			0.14	0.52	0.20	0.16	0.50	0.22	0.43	0.12	0.49	0.18	**-0.07**	**-0.03**	0.14
				0.28	0.52	0.37	0.30	0.28	0.41	**-0.01**	0.16	0.16	0.26	0.30	0.39
					0.33	0.30	0.68	0.39	0.56	0.45	0.16	0.43	0.03	0.30	0.39
						0.50	0.43	0.52	0.50	0.09	0.37	0.07	0.47	0.35	0.37
							0.28	0.49	0.28	0.16	0.30	0.45	0.47	0.20	0.41
								0.41	0.50	0.43	0.07	0.33	0.05	0.09	0.49
									0.22	0.03	0.39	0.35	0.56	0.41	0.47
										0.31	0.30	0.18	0.12	0.31	0.14
											-0.24	0.33	**-0.10**	**-0.07**	0.33
												0.05	0.37	0.18	0.09
													0.22	0.07	0.35
														0.39	0.33
															0.18

Table 3: Kendall tau correlation between each pair of ranked lists. The first fifteen columns correspond to the pairwise rankings from social media professionals, shown in Table 2. The final column corresponds to the ranking from the social media test, shown in Table 1. Negative values are highlighted in bold.

Kendall	Spearman
2.9	4.1
6.6	8.6
6.5	8.7
3.8	5.5
4.3	6.2
6.4	8.4
5.9	7.7
5.5	7.4
5.8	7.7
5.6	7.6
5.7	7.8
2.7	3.7
3.0	4.3
4.1	5.6
3.7	5.1
3.6	5.2
5.2	7.2

Table 4: Left column: ith row gives the sum of the pairwise Kendall tau coefficients from Table 3 involving list i. The final row corresponds to the list produced by the social media test. Right column: same results using the Spearman rho correlation coefficient. Here, a larger number indicates a greater degree of consistency with the other, independent, views.

publication. PG was supported by the Research Councils UK Digital Economy Programme via EPSRC grant EP/G065802/1 *The Horizon Digital Economy Hub*. DJH acknowledges support from a Royal Society Wolfson Award and an EPSRC/Digital Economy Established Career Fellowship grant EP/M00158X/1. JAW was supported by EPSRC grant EP/I016058/1.

References

[1] S. ARAL, *Social science: Poked to vote*, Nature, 489 (2012), pp. 212–214.

[2] S. ARAL AND D. WALKER, *Identifying influential and susceptible members of social networks*, Science, 337 (2012), pp. 337–341.

[3] E. BAKSHY, J. M. HOFMAN, W. A. MASON, AND D. J. WATTS, *Everyone's an influencer: quantifying influence on Twitter*, in Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, New York, NY, USA, 2011, ACM, pp. 65–74.

[4] E. BAKSHY, I. ROSENN, C. MARLOW, AND L. ADAMIC, *The role of social networks in information diffusion*, in Proceedings of the 21st international conference on World Wide Web, WWW '12, New York, NY, USA, 2012, ACM, pp. 519–528.

[5] A. BOUTET, H. KIM, AND E. YONEKI, *Whats in Twitter, I know what parties are popular and who you are supporting now!*, Social Network Analysis and Mining, 3 (2013), pp. 1379–1391.

[6] R. J. BRODIE, A. ILIC, B. JURIC, AND L. HOLLEBEEK, *Consumer engagement in a virtual brand community: An exploratory analysis*, Journal of Business Research, 66 (2013), pp. 105–114.

[7] K. W. CHANA AND S. Y. LIB, *Understanding consumer-to-consumer interactions in virtual communities: The salience of reciprocity*, Journal of Business Research, 63 (2010), pp. 1033–1040.

[8] J. CHUA, M. ARCE-URRIZAB, J.-J. CEBOLLADA-CALVOC, AND P. K. CHINTAGUNTAD, *An empirical analysis of shopping behavior across online and offline channels for grocery products: The moderating effects of household and product characteristics*, Journal of Interactive Marketing, 24 (2010), pp. 251–268.

[9] F. CIULLA, D. MOCANU, A. BARONCHELLI, B. GONÇALVES, N. PERRA, AND A. VESPIGNANI, *Beating the news using social media: the case study of American Idol*, EPJ Data Science, 1 (2012), pp. 1–11.

[10] P. J. DANAHER, I. W. WILSON, AND R. A. DAVIS, *A comparison of online and offline consumer brand loyalty*, Marketing Science, 22 (2003), pp. 461–476.

[11] P. FARHI, *Oreo's tweeted ad was Super Bowl blackout's big winner*, Washington Post, (February 05, 2013).

[12] S. L. FELD, *Why your friends have more friends than you do*, American Journal of Sociology, 96 (1991), pp. 1464–1477.

[13] M. GARCÍA-HERRANZ, E. MORO, M. CEBRIÁN, N. A. CHRISTAKIS, AND J. H. FOWLER, *Using friends as sensors to detect global-scale contagious outbreaks*, PLOS ONE, 9(4) (2014), p. e92413.

[14] P. GRINDROD, *Mathematical Underpinnings of Analytics: Theory and Applications*, Oxford University Press, Oxford, 2014.

[15] C. KUEHN, E. A. MARTENS, AND D. M. ROMERO, *Critical transitions in social network activity*, Journal of Complex Networks, 2 (2014), pp. 141–152.

[16] H. KWAK, C. LEE, H. PARK, AND S. MOON, *What is Twitter, a social network or a news media?*, in Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA, 2010, ACM, pp. 591–600.

[17] P. LAFLIN, A. V. MANTZARIS, F. AINLEY, A. OTLEY, P. GRINDROD, AND D. J. HIGHAM, *Discovering and validating influence in a dynamic online social network*, Social Network Analysis and Mining, 3 (2013), pp. 1311–1323.

[18] ——, *Anticipating activity in social media spikes*, in Proceedings of the Workshop on Modelling and Mining Temporal Interactions Workshop of the 9th International Conference on the Web and Social Media, Oxford, CA, USA, 2015, Association for the Association for the Advancement of Artificial Intelligence.

[19] M. LAROCHE, M. R. HABIBI, AND M.-O. RICHARD, *To be or not to be in social media: How brand loyalty is affected by social media?*, International Journal of Information Management, 33 (2013), pp. 76–82.

[20] M. LAROCHE, M. R. HABIBI, M.-O. RICHARD, AND R. SANKARANARAYANAN, *The effects of social media based brand communities on brand community markers, value creation practices, brand trust and brand loyalty*, Computers in Human Behavior, 28 (2012), pp. 1755–1767.

[21] D. LAZER, A. PENTLAND, L. ADAMIC, S. ARAL, A.-L. BARABÁSI, D. BREWER, N. CHRISTAKIS, N. CONTRACTOR, J. FOWLER, M. GUTMANN, AND T. JEBARA, *Computational social science*, Science, 323 (2009), pp. 721–723.

[22] S. H. LEE, P.-J. KIM, AND H. JEONG, *Statistical properties of sampled networks*, Phys. Rev. E, 73 (2006), p. 016102.

[23] J. LESKOVEC AND C. FALOUTSOS, *Sampling from large graphs*, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06, New York, NY, USA, 2006, ACM, pp. 631–636.

[24] C. LOWCAY, S. MARSLAND, AND C. MCCARTIN, *Network parameters and heuristics in practice: a case study using the target set selection problem*, Journal of Complex Networks, 2 (2014), pp. 373–393.

[25] A. S. MAIYA AND T. Y. BERGER-WOLF, *Benefits of bias: Towards better characterization of network sampling*, in In Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11), San Diego, 2011.

[26] M. STUMPF, C. WIUF, AND R. MAY, *Subnets of scale-free networks are not scale-free: Sampling properties of networks*, Proc. Nat. Acad. Sci., 102 (2005), pp. 4221–4224.

[27] S. WU, J. M. HOFMAN, W. A. MASON, AND D. J. WATTS, *Who says what to whom on Twitter*, in Proceedings of the 20th international conference on World wide web, WWW '11, New York, NY, USA, 2011, ACM, pp. 705–714.

[28] M. E. ZAGLIA, *Brand communities embedded in social networks*, Journal of Business Research, 66 (2013), pp. 216–223.